Multi-level Syntactic Representation in the Szeged FC Treebank

Katalin Ilona Simkó[1], Veronika Vincze[2], Richárd Farkas[1]

1 University of Szeged, Department of Informatics
kata.simko@gmail.com
rfarkas@inf.u-szeged.hu
[2] MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

The two most widely used syntactic theories among the existing ones are constituent and dependency syntactic theories. The Szeged Treebank contains manually annotated syntactic trees in both constituent and dependency formats. Both analyses have their advantages and disadvantages as well. The constituent representation groups words that are part of the same unit of meaning into phrases, while dependency grammars connect the words of the sentence directly to each other without the use of abstract nodes.

It is undecided whether either of these grammars can be considered superior for the analysis of Hungarian and other morphologically rich languages, as both representations contain important information on their syntax. We have therefore decided to create a syntactic representation in which the information encoded in both of these structures is preserved.

In order to make use of the benefits of both, we are currently working on a complex syntactic representation for the sentences of the Szeged Treebank that utilizes the constituent and the dependency trees as well as the morphological analysis of the words. The new structure analyses different types of syntactic information at different levels, similar to Lexical-Functional Grammar. This multi-level syntactic representation is created by automatic conversion of the already existing constituent and dependency trees and the words' morphological analyses available for the sentences of the Szeged Treebank.

The phrase structures of the constituent analysis are represented here in a c-structure reflecting the surface structure of the sentences. These are converted directly from the constituent trees of the Szeged Treebank.

The sentences' argument structure is represented at a different level, in the f-structure. We convert these using the dependency trees and the morphological information on the words of the sentence.

The new database enables the training and evaluation of statistical syntactic parsers with a new approach, as well as testing these in real-world natural language processing tasks. Thus the usefulness of this multi-level syntactic representation can be empirically compared to that of the classical constituent and dependency analyses as well.